EVROS: All-Optical Programmable Disaggregated Data Centre Interconnect Utilizing Hollow-Core Bandgap Fibre

G. M. Saridis[(1)], Y. Yan[(1)], Y.Shu[(1)], S. Yan[(1)], M. Arslan[(1)], T. Bradley[(2)], N. V. Wheeler[(2)], N.H.L. Wong[(2)], F. Poletti[(2)], M.N. Petrovich[(2)], D.J. Richardson[(2)], S. Poole[(3)], G. Zervas[(1)], D. Simeonidou[(1)]

[(1)] High Performance Networks, University of Bristol, United Kingdom, (George.Saridis@bristol.ac.uk)
[(2)] Optoelectronics Research Centre, University of Southampton, United Kingdom
[(3)] Finisar Australia, Rosebery, Australia

Abstract *We present an all-optical flexible disaggregated flat DCN architecture utilizing Hollow-Core Bandgap Fibre, reconfigurable 4×16/8×12 Spectrum Selective Switches and FPGA-based switch and interface intra/inter-blade cards enabling chip-level access while utilizing cut-through low-latency tuneable WDM/WDM-TDM.*

Introduction

The growth of cloud services, HPC applications, big data storage/processing along with multi-tenancy, imposes new requirements that challenge current data center networks (DCN)[1]. Legacy multi-layer DCNs will have detrimental effects on the distributed and highly-parallelized computing architectures[2]. They lack modularity and flexibility in resource allocation and introduce delays and significant power consumption due to the frequent O/E/O conversions. Optical Top of Rack (ToR) to ToR DCN architectures have been proposed using WDM[3-4] and/or SDM/TDM[5], showing higher capacity, lower latency and improved scalability. A fully disaggregated DCN with 25/75 local to remote memory balance can be realized with (sub)-μsec latency[6-7]. Apart from low-latency, a disaggregated architecture should be modular (data center in a box), flexible, programmable and scalable to handle diverse services and requirements.

In this paper, we propose and experimentally demonstrate, for the first time, a programmable, rapidly-tuneable and re-configurable disaggregated flat DCN architecture with all-optical chip-to-chip inter-blade interconnection within and between racks. It delivers ultra-low deterministic latency between blade E/O-I/Os within (44.8–134.4 ns) and between (544 ns) racks, with an additional 316 ns for the FPGA-based electronic blade E/O-I/O from/to memory direct memory access (DMA) driver. Granularity of 100 Mb/s to 142 Gb/s per blade to serve mice and elephant flows is supported together with up to 1-to-77 (# λ and slots) connectivity. Hollow-core photonic bandgap fibre (HC-PBGF)[8] links offer 30% propagation delay reduction and dimension-programmable single device 20-port NxM spectrum selective switches (SSS) were used for ToR and top of cluster (ToC) switches. An intra/inter-blade programmable FPGA-based Switch & Interface Card (SIC), which performs Remote DMA and facilitates 2 hybrid 10G OOK fast-tunable ports (to avoid DSP delays) supporting non-disruptive switch-over between fast Ethernet-Lite WDM and WDM/TDM for front-end and 2×10G OOK cut-through back-end ports, allows for pure intra-rack and front-end (hybrid Intra/Inter Rack) networking respectively.

Flexible disaggregated DCN architecture

Fig. 1 shows the proposed architecture with the bidirectional intra-cluster interconnection scheme of various blades/racks containing disaggregated resources for increased system modularity and performance. Two complementary physical networks co-exist, the back-end (intra-rack) and the front-end (intra/inter rack) (Fig. 1 inset A) to support diverse interconnection scenarios. The back-end full-mesh intra-rack FPGA-based network provides all-to-all (zero-to-multi-hop) communication. The front-end re-configurable

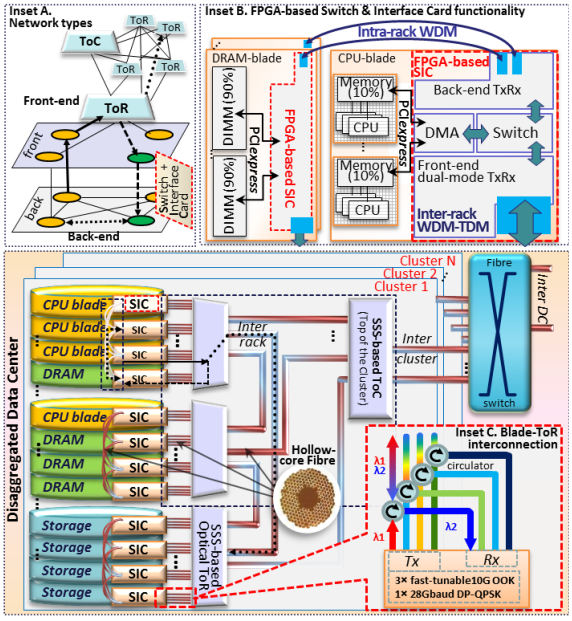

Fig. 1: Overall proposed DCN architecture (bottom) Inset A: Back- & Front-end description. Inset B: FPGA-based SIC. Inset C: Blade-ToR interconnection

intra/inter-rack network provides one, two or three hop intra-cluster communication. The optical SSS-based ToR switches, which are bandwidth-to-port selection-flexible and able to dynamically reconfigure their I/O dimensions (e.g. 4×16, 8×12, 10×10), groom, switch and balance traffic within and between racks. This unique feature enables the architecture to adapt to different networking scenarios with diverse requirements on capacity, latency and connectivity. The identical SSS-based ToC handles and re-balances intra/inter-cluster communication and aggregates traffic towards a high-port count fibre switch, adding scalability to the network. The blade-to-ToR interface is implemented with 4 channels (Fig.1 inset C) and circulators to exploit the bi-directionality of the SSS, thus saving resources. By utilizing 3×fast tuneable 10 Gb/s OOK channels for either low-latency WDM or WDM/TDM and a 28 Gbaud DP-QPSK channel on each SIC, wide (C-band), flexible-bandwidth connectivity with various granularities is achieved.

The highly programmable FPGA-based SIC (Fig.1 inset B) is developed to read/write data blocks via a PCIe Gen2 bus between the CPU memory and remote DRAM chip. Then accordingly, traffic is switched either via the cut-through optically-connected electrically-switched back-end or via the programmable fast-tuneable dual-mode (cut-through WDM/WDM-TDM) to reach remote destinations. Apart from interfacing, SIC can also hitlessly switch traffic from the back-end towards the front-end and vice versa, allowing multi-hop communication to assist congested links. Furthermore, SIC is able to flexibly aggregate TDM traffic based on network requirements (QoS, latency, connectivity) by dynamically programming the size/number of slots per frame.

Experimental setup, results and evaluation

The experimental test-bed (Fig. 3), contains two FPGA-based SIC boards that support real-traffic scenarios with fully-functional transceivers that act as the interface between the local memory chip and the back/front-end network, three re-configurable 4×16/8×12 SSSs, a traffic analyser and nine spools of HC-PBGF (6×10 m for intra-rack and 3×100 m for inter-rack). We demonstrate 85 channels in total (77 fast-tuneable + 8 ECL) with variable channel-spacing and baud-rates (Fig. 2 top-left). The maximum capacity-per-blade is 142G (i.e. 3×10G + 1×112G) and per-rack 568G (i.e. 4×blades x 142G). Diverse and variable levels of granularity are supported per-single-carrier and per-blade, ranging from 100 Mb/s to 7.8 Gb/s (fast WDM-TDM). Transmission-wise, 8 channels (from the ECL bank) are placed between the 50 GHz-spaced GCSR fast-tuneable channels, creating a 25 GHz channel spacing spectral slice which feeds the 10G MZM modulator, driven by the FPGA-based SIC. In the middle stage, signals pass through circulators, 10/100 m HC-PBGF links, 1/2/3 SSS and terminate to another FPGA as seen in Fig. 2. The BER is tested with real traffic (with PRBS payload) to evaluate board-to-board scenarios (optics & electronics) in Fig. 3a. In the 25(50) GHz-spaced 10G signals, 17(32) dB OSNR is observed and ≤2 dB penalty was obtained for intra-rack interconnection with a further ≤3 dB penalty for the two and three-hop inter-rack cases, respectively.

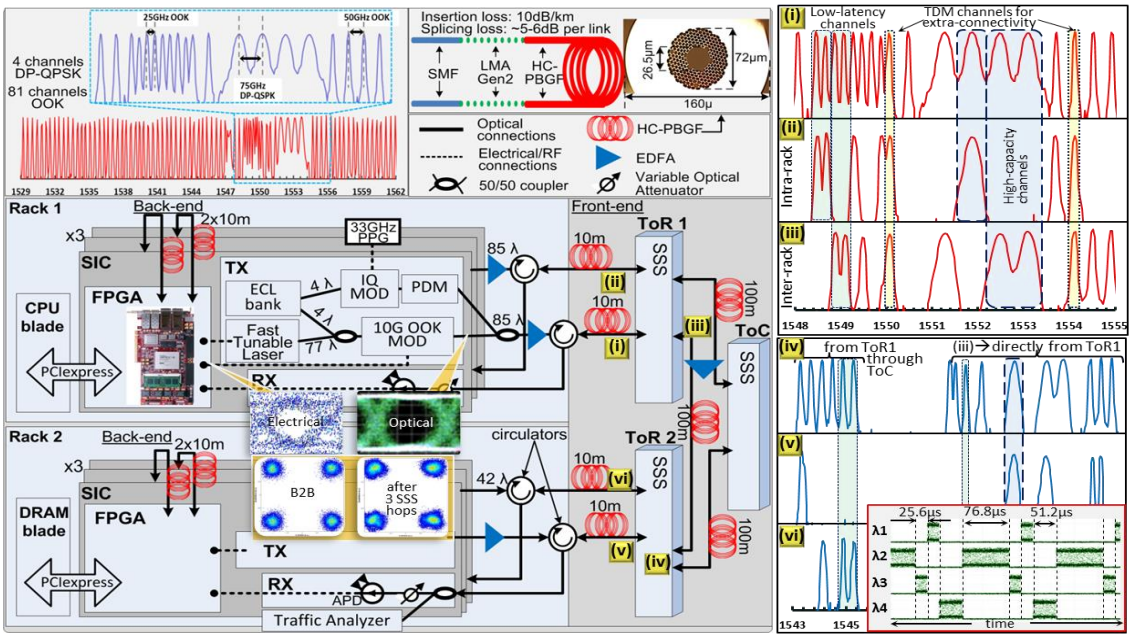

Fig. 2: Experimental setup, HC-PBGF characteristics, spectra at key points of the network, time plots

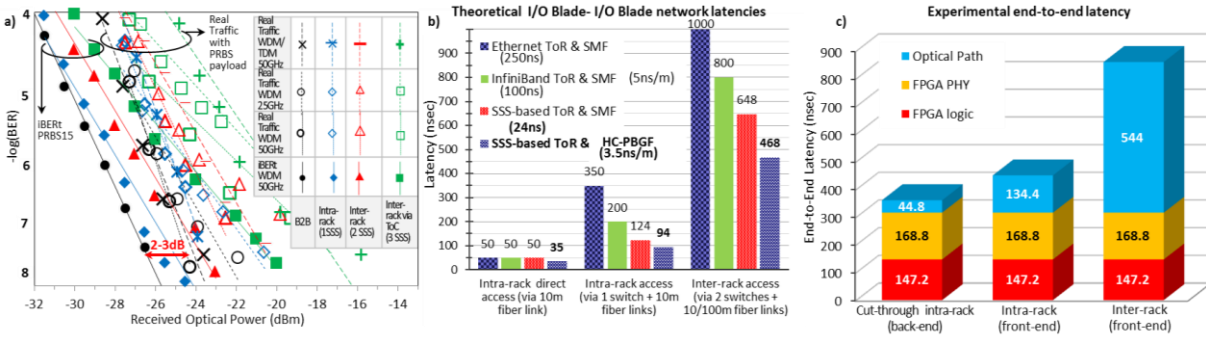

Fig. 3: a) BER curves b) Theoretical network-only latency c) Measured blade-to-blade latency – excluding DMA overhead

By configuring all dynamic and controlled elements of the SIC in conjunction with the SSSs in various frequency-to-port combinations and by changing between the 4×16/8×12 arrangements, we demonstrate the scenarios of the proposed architecture, as seen in the spectra and time plots of Fig. 2. In Fig. 2, different WDM channels entering ToR1 *(i)* in Rack1, are diverted towards either intra-*(ii)* or inter-rack *(iii)* ports. Then in ToR2 *(iv)*, channels that arrived from ToR1 directly or via ToC, are switched to two separate blades *(v)*, *(vi)* of Rack 2 (or vice versa due to bi-directionality). In the time domain (inset in Fig. 2), λ1-4 start from point *(i)*, are modulated in different time-slots (25.6/51.2/76.8μs etc.) in order to deliver data in different destinations according to the SSS configuration. λ1 stays intra-rack via ToR1, λ2-3 go directly inter-rack to two different blades of ToR2, whereas λ4 goes to ToR2 via the ToC for congestion avoidance, due to blocking in the direct link between ToR1 and ToR2.

In Fig. 3b, two all-optical (SSS+SMF/HC-PBGF) interconnection schemes are theoretically compared with two electronic (Mellanox Ethernet/Infiniband ToR & SMF) according to their typical values of latency (network only), showing the advantage of our proposed architecture. In Fig. 3c, end-to-end latency is measured from memory chip-to-chip via DMA, including FPGA PHY, logic and optical network delays. The DMA driver latency (1.5 μsec) is excluded due to server, OS and driver dependencies, showing total 360.8, 450.4, 860 nsec latency for cut-through intra-rack, intra- and inter-rack via front-end respectively.

Conclusions

We have successfully demonstrated an all-optical chip-to-chip inter-blade and inter-rack interconnect for highly- connected flexible communication between disaggregated resources (processing/memory/storage) for modern modular HPC/Data Centres. By utilizing FPGAs for real-traffic chip-scale memory access and switching, fast-tuneable WDM/TDM transceivers, re-configurable flexi-grid optical SSS switches along with HC-PBGF, the proposed network offers variable capacity and granularity (from 100 Mb/s to 158 Gb/s per blade), high-spectral efficiency, high-connectivity (1-to-77 per port) and ultra-low latency interconnection (360.8 nsec intra-rack & 860 nsec inter-rack), programmable to support diverse and unpredictable Data Center services.

Acknowledgements

This work is supported by EPSRC EP/I01196X: The Photonics Hyperhighway, EP/L027070/1 SONATAS and ECFP7 grant no. 619572, COSIGN. The authors are also grateful to Yenista for providing their Optical Spectrum Analyzer (OSA20) for the experiments.

References

[1] R. Branch et al., "Cloud Computing and Big Data: A Review of Current Service Models and Hardware Perspectives," Journal of Software Engineering and Applications, vol. **7**, no. 8, pp. 686–693, (2014).

[2] J. Diaz et al., "A Survey of Parallel Programming Models and Tools in the Multi and Many-Core Era," IEEE Trans. on Parallel and Distributed Systems, vol. **23**, no. 8, pp. 1369–1386, (2012).

[3] Z. Cao et al., "Experimental demonstration of dynamic flexible bandwidth optical data center network with all-to-all interconnectivity," Proc. ECOC, PDP 1.1, (2014).

[4] G. Saridis et al., "DORIOS: Demonstration of an All-Optical Distributed CPU, Memory, Storage Intra DCN Interconnect," Proc. OFC, W1D.2, (2015).

[5] S. Yan et al., "First Demonstration of All-Optical Programmable SDM/TDM Intra Data Centre and WDM Inter-DCN Communication," Proc. ECOC, PDP. 1.2. (2014).

[6] S. Han et al., "Network support for resource disaggregation in next-generation datacenters," Proc. ACM Workshop on Hot Topics in Networks, p. 10, (2013).

[7] K. Lim et al., "System-level implications of disaggregated memory," IEEE HPCA, pp. 1–12, (2012).

[8] F. Poletti et al., "Towards high-capacity fibre-optic communications at the speed of light in vacuum," Nature Photon., vol. **7**, no. 4, pp. 279–284, (2013).